Differentiable Probabilistic Models

William Watson
nextbillyonair@gmail.com

Abstract

d

Contents

1 Introduction

1.1 Philosophy

2 Background

2.1 Kronecker Product

2.2 Gradients

2.3 Jacobian

2.4 Hessian

2.5 Newton Optimization

3 Distributions

3.1 Distribution (Base Class)

3.2 Arcsine

3.3 Asymmetric Laplace

3.4 Bernoulli

3.5 Beta

3.6 Categorical

3.7 Cauchy

3.8 Chi Square

3.9 Conditional Model

3.10 Convolution

3.11 Data

3.12 Dirac Delta

3.13 Dirichlet

3.14 Exponential

3.15 Fisher-Snedcor (F-Distribution)

3.16 Gamma

3.17 Generator

3.18 Gumbel

3.19 Gumbel Softmax

3.20 Gumbel Mixture Model

3.21 Half Cauchy

3.22 Half Normal

3.23 Hyperbolic Secant

3.24 Infinite Mixture Model

3.25 Kumaraswamy

3.26 Langevin

3.27 Laplace

4.1 Transform (Base Class)

4.2 Inverse Transform

4.3 Chain

4.4 Affine

- **Parameters**
  - Location $\mu \in \mathbb{R}^n$
  - Scale $\sigma > 0$
- **Forward**
$$f(x) = \mu + \sigma \cdot x \tag{1}$$
- **Inverse**
$$f^{-1}(y) = \frac{y - \mu}{\sigma} \tag{2}$$
- **Log Absolute Determinant Jacobian**
$$\log |\det \mathbf{J}|(x, y) = \log |\sigma| \tag{3}$$

4.5 Exp

- **Parameters**
  - None
- **Forward**
$$f(x) = e^x \tag{4}$$
- **Inverse**
$$f^{-1}(y) = \log y \tag{5}$$
- **Log Absolute Determinant Jacobian**
$$\log |\det \mathbf{J}|(x, y) = x \tag{6}$$

4.6 Expm1

- **Parameters**
  - None
- **Forward**
$$f(x) = e^x - 1 \tag{7}$$
- **Inverse**
$$f^{-1}(y) = \log(1 + y) \tag{8}$$
- **Log Absolute Determinant Jacobian**
$$\log |\det \mathbf{J}|(x, y) = x \tag{9}$$

4.7 Gumbel

- **Parameters**
  - Location $\mu \in \mathbb{R}^n$
  - Scale $\sigma > 0$
- **Forward**
$$f(x) = \exp\left(-\exp\left(-\frac{x - \mu}{\sigma}\right)\right) \tag{10}$$
- **Inverse**
$$f^{-1}(y) = \mu - \sigma \cdot \log\left(-\log\left(y\right)\right) \tag{11}$$
- **Log Absolute Determinant Jacobian**
$$\log |\det \mathbf{J}|(x, y) = -\log\left(\frac{\sigma}{-\log(y) \cdot y}\right) \tag{12}$$

4.8 Identity

4.9 Kumaraswamy

4.10 Log

- **Parameters**
  - None
- **Forward**

$$f(x) = \log x \tag{13}$$

- **Inverse**

$$f^{-1}(y) = \exp y \tag{14}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = -y \tag{15}$$

4.11 Logit

- **Parameters**
  - None
- **Forward**

$$f(x) = \log\left(\frac{x}{1-x}\right) \tag{16}$$

- **Inverse**

$$f^{-1}(y) = \frac{1}{1+e^{-y}} \tag{17}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = \log\left(1+e^{-y}\right) + \log\left(1+e^{y}\right) \tag{18}$$

4.12 Planar

4.13 Power

- **Parameters**
  - Power p
- **Forward**

$$f(x) = \begin{cases} e^x & p = 0 \\ (1 + x \cdot p)^{1/p} & \text{otherwise} \end{cases} \tag{19}$$

- **Inverse**

$$f^{-1}(y) = \begin{cases} \log y & p = 0 \\ y^{p-1}/p & \text{otherwise} \end{cases} \tag{20}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = \begin{cases} x & p = 0 \\ \left(\frac{1}{p} - 1\right) \cdot \log(x \cdot p + 1) & \text{otherwise} \end{cases} \tag{21}$$

4.14 Radial

4.15 Reciprocal

- **Parameters**
  - None
- **Forward**

$$f(x) = 1/x \tag{22}$$

- **Inverse**

$$f^{-1}(y) = 1/y \tag{23}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = -2 \cdot \log|x| \tag{24}$$

4.16 Sigmoid

- **Parameters**
  - None
- **Forward**

$$f(x) = \frac{1}{1+e^{-x}} \tag{25}$$

- **Inverse**

$$f^{-1}(y) = \log\left(\frac{y}{1-y}\right) \tag{26}$$

- **Log Absolute Determinant Jacobian**

$$\log|\det \mathbf{J}|(x, y) = -\log\left(1+e^{-x}\right) - \log\left(1+e^{x}\right) \tag{27}$$

4.17 SinhArcsinh

4.18 Softplus

4.19 Softsign

4.20 Square

4.21 Tanh

4.22 Weibull

5 Criterion

The criterion and divergences listed here can be used to quantify the "distance" between two distributions. Hence, in conjunction with torch optimizers, one can minimize said difference to learn the paramters of a distribution. For sake of notation clarity, p is the true distribution and q is the learned distribution. Hence we "fit" q to match p. In addition, we provide the Monte Carlo approximation.

		P		Q	
	Criterion	$\log p(x)$	$x \sim P$	$\log q(x)$	$x \sim Q$
Divergence	Cross-Entropy		✓	✓	
	Perplexity		✓	✓	
	Exponential	✓	✓	✓	
	Forward KL	✓	✓	✓	
	Reverse KL	✓		✓	✓
	JS Divergence	✓	✓	✓	✓
Adversarial	GAN		✓		✓
	MMGAN		✓		✓
	WGAN		✓		✓
	LSGAN		✓		✓

5.1 Divergences

5.1.1 Cross-Entropy

$$\begin{aligned} H(p,q) &= -\int p(x)\log q(x)dx \\ &= -\frac{1}{n}\sum_{x\sim p}\log q(x) \end{aligned} \tag{28}$$

5.1.2 Perplexity

$$\begin{aligned} H(p,q) &= \exp\left(-\int p(x)\log q(x)dx\right) \\ &= \exp\left(-\frac{1}{n}\sum_{x\sim p}\log q(x)\right) \end{aligned} \tag{29}$$

5.1.3 Exponential

5.1.4 Forward KL Divergence

$$\begin{aligned} H(p,q) &= \int p(x)\log\frac{p(x)}{q(x)}dx \\ &= \frac{1}{n}\sum_{x\sim p}\log\frac{p(x)}{q(x)} \end{aligned} \tag{30}$$

5.1.5 Reverse KL Divergence

$$\begin{aligned} H(p,q) &= \int q(x)\log\frac{q(x)}{p(x)}dx \\ &= \frac{1}{n}\sum_{x\sim q}\log\frac{q(x)}{p(x)} \end{aligned} \tag{31}$$

5.1.6 Jensen-Shannon Divergence

5.1.7 Earth Mover's Distance

5.2 Adversarial Loss

Adversarial Losses are criterion functions that allow for sample-sample based training between models p and q. More formally, it hides a Discriminator model that attempts to discriminate between the real data from p and fake data generated from q.

5.2.1 Adversarial Loss (Base Class)

5.2.2 GAN Loss

5.2.3 MMGAN Loss

5.2.4 WGAN Loss

5.2.5 LSGAN Loss

5.2.6 Gradient Penalty

5.2.7 Spectral Norm

5.3 ELBO

6 Models

6.1 Regression

6.1.1 Linear Regression (Normal)

6.1.2 L1 Regression (Laplace)

6.1.3 Ridge Regression (Normal + Normal Prior on Weights)

6.1.4 Lasso Regression (Normal + Laplace Prior on Weights)

6.2 Classification

6.2.1 Logistic Regression (Bernoulli)

6.2.2 Bayesian Logistic Regression (Bernoulli)

6.2.3 Softmax Regression (Categorical)

6.2.4 Bernoulli Naive Bayes (Bernoulli - Bernoulli)

6.2.5 Gaussian Naive Bayes (Multinomial - Gaussian)

6.2.6 Multinomial Naive Bayes (Bernoulli - Multinomial)

6.2.7 Linear Discriminant Analysis (Multinomial - Shared Covariance)

6.2.8 Quadratic Discriminant Analysis (Multinomial - Multivariate Gaussian)

6.3 Clustering

6.3.1 Gaussian Mixture Model

6.4 Unconstrained Matrix Factorization (Gaussian)

6.5 Principle Components Analysis

6.5.1 PCA

6.5.2 EM-PPCA

6.5.3 Variational PPCA

6.6 Generative Adversarial Networks

6.6.1 Generative Adversarial Networks

6.6.2 GAN Model

6.6.3 MMGAN Model

6.6.4 WGAN Model

6.6.5 LSGAN Model

6.7 Variational Auto-Encoders

7 Monte Carlo